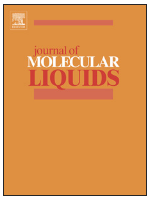
Journal of Molecular Liquids xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Molecular Liquids

journal homepage: www.elsevier.com/locate/molliq



Prediction of gas-to-ionic liquid partition coefficient of organic solutes dissolved in 1-(2-methoxyethyl)-1-methylpyrrolidinium *tris*(pentafluoroethyl)trifluorophosphate using QSPR approaches

Q1 Zahra Dashtbozorgi [a,*], Hassan Golmohammadi [b], William E. Acree Jr. [c]

[a] Young Researchers and Elite Club, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[b] Young Researchers and Elite Club, Shahr-e-Rey Branch, Islamic Azad University, Tehran, Iran
[c] Department of Chemistry, P. O. Box 305070, University of North Texas, Denton, TX 76203-5070, USA

ARTICLE INFO

Article history:
Received 22 September 2014
Received in revised form 17 November 2014
Accepted 19 November 2014
Available online xxxx

Keywords:
Quantitative structure–property relationship
Gas-to-ionic liquid partition coefficient
Replacement method
Support vector machine

ABSTRACT

The present work represents a quantitative structure–property relationship (QSPR) study for predicting the gas-to-ionic liquid partition coefficients (log *K*) of organic solutes dissolved in 1-(2-methoxyethyl)-1-methylpyrrolidinium *tris*(pentafluoroethyl)trifluorophosphate based on replacement method (RM) and support vector machine (SVM). The variable selection technique of replacement method (RM) was employed to select the most favorable subset of descriptors from the more than 1000 theoretical descriptors calculated using the Dragon package. The six descriptors selected were used as inputs of SVM to predict the gas-to-ionic liquid partition coefficient of a set of 92 aliphatic and aromatic solutes dissolved in 1-(2-methoxyethyl)-1-methylpyrrolidinium *tris*(pentafluoroethyl)trifluorophosphate. Results of our computations demonstrate that SVM can be used as a substitute powerful modeling tool for QSPR studies. The log *K* values calculated by SVM were in good agreement with the experimental data, and the performances of the SVM models were superior to RM one.

1. Introduction

Ionic liquids (ILs) usually refer to organic salts with comparatively low melting point temperatures (below 100 °C) [1]. Ionic liquids exhibit unique characteristics such as extremely low vapor pressure, good thermal stability, electrical conductivity, and high polarity. The miscibility of ionic liquids with water or organic solvents depends on the cationic–anionic combination, and the functional groups and lengths of the alkyl-chains on the cation. An extensive range of applications using ionic liquids have been reported in many areas such as catalysis, organic chemistry, electrochemistry, and separation science [2–8].

ILs can dissolve a wide range of organic, organometallic, and inorganic compounds [9,10], and as stated above ILs have negligible vapor pressures. There is little (if any) loss of solvent through evaporation with ionic liquids. This avoids the environmental and safety concerns that result from solvent volatilization, as is the case in traditional organic solvents. Ionic liquids have been recommended as a possible substitute for the more traditional organic solvents that are often toxic, highly flammable, and volatile. ILs have the potential to be alternative reaction media for “Green Chemistry” [11,12].

Analytical methods for ionic liquid characterization are challenging owing to the complexity of the cationic or anionic organic ions, counter-ions, and ionic impurities. Ion chromatography (IC), liquid chromatography (LC), and hydrophilic interaction liquid chromatography (HILIC) have been used for ionic liquid analysis, featuring ion-exchange or reversed phase columns [13,7,8].

The thermodynamic gas-to-ionic liquid partition coefficient, *K*, can be computed from isothermal gas–liquid chromatographic measurements through:

$$K = V_N/V_L \qquad (1)$$

where V_N is the volume of the carrier gas needed to elute the solute, and Q3 V_L is the volume of liquid present as the stationary phase [14]. Physical and thermodynamic property data of organic compounds such as partition coefficient are significant in the engineering design and operation of industrial chemical processes. Since the experimental determination of gas-to-ionic liquid partition coefficient is time-consuming and expensive, and there is increased require of reliable physical and thermodynamic data for the optimization of chemical processes, it would Q4 be very useful to develop predictive models that can be used to predict these properties of organic compounds that are not synthesized or their properties are unknown.

Alternatively, the quantitative structure–property relationship (QSPR) provides a talented method for the estimation of the partition coefficient of organic compounds based on descriptors derived solely from the molecular structure to fit experimental data [15]. The QSPR

[*] Corresponding author.
E-mail address: z.dashtbozorgi@gmail.com (Z. Dashtbozorgi).

http://dx.doi.org/10.1016/j.molliq.2014.11.025
0167-7322/© 2014 Published by Elsevier B.V.

approach has become very practical in the prediction of physical and chemical properties [16]. The support vector machine (SVM) was developed from the machine learning community by Vapnik and co-workers in 1995 [17,18]. It is a new algorithm developed for regression and classification, and has shown good performance in classification problems through several successful applications [19–25]. SVM has also demonstrated good performance in QSPR studies due to its aptitude for interpreting the nonlinear relationships between molecular structure and properties [26–34].

In the present study, SVM was performed for the first time to describe the gas-to-ionic liquid partition coefficient (log *K*) of 92 organic solutes dissolved in a 1-(2-methoxyethyl)-1-methylpyrrolidinium *tris*(pentafluoroethyl)trifluorophosphate, ($[\text{MeoeMPyrr}]^+[\text{FAP}]^-$), at 323 K. The main aim of the present study was to generate a QSPR model that could be used for the prediction of log *K* of a diverse set of compounds from solely molecular structure considerations. A secondary goal was to demonstrate the flexible modeling ability of SVM. The Replacement method (RM) was also employed to construct quantitative linear relationship to compare with the results obtained by SVM.

2. Methodology

2.1. Data set

The experimental data set of gas-to-ionic liquid partition coefficients of 92 organic solutes dissolved in ($[\text{MeoeMPyrr}]^+[\text{FAP}]^-$) extracted from the values reported by Jiang et al. [35]. The molecules in data set contained alkanes, alkenes, alkynes, alkyl halides, alcohols, phenols, ethers, esters, ketones, aldehydes, amines, anilines, nitriles, nitro compounds, polycyclic hydrocarbons, heterocyclic compounds, benzene derivatives, etc., are summarized in Table 1. The partition coefficients of all molecules included in data set were obtained under nearly identical experimental conditions and refer to a temperature of 323 K. The partition coefficients fall in the range of $\log K = 0.816$ to $\log K = 4.721$ for pentane and naphthalene, respectively. The entire dataset was arbitrarily divided into two subsets. A training set of 61 compounds and a prediction set of 31 compounds. The training set was used to build and optimize the QSPR model and the external prediction set was used to assess the prediction power of the obtained model.

2.2. Molecular modeling and descriptor calculation

A main step in each QSPR study is selecting and calculating the structural descriptors as numerical encoded parameters representing the chemical structures. The generated numerical descriptors were responsible for encoding important features of the structures. Owing to the variety of the molecules studied, different kinds of descriptors were calculated. The calculation process of the molecular descriptors is described as follows: In the first step, all structures were drawn with the HyperChem (Ver. 7.0) [36] and then pre-optimized using MM + molecular mechanics force field. All calculations were performed at the restricted Hartree–Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root-mean-square gradient was 0.001. In a next step, the Hyperchem output files were used by the Dragon package (Version 3) to calculate molecular descriptors [37]. More than 1400 theoretical descriptors were calculated regularly for each molecule by this software. The calculated descriptors can be categorized into several groups, 0D, constitutional descriptors; 1D, functional groups, atom-centered fragments, empirical descriptors and molecular properties; 2D, topological descriptors, molecular walk counts, BCUTs descriptors, Galvez topological charge indices, 2D autocorrelations; 3D, aromaticity indices, Randic molecular profiles from the geometry matrix, geometrical, RDF, 3D-MORSE, WHIMs, and GETAWAYs descriptors.

The calculated descriptors were first analyzed for the existence of constant or near constant variables. The detected ones were then removed. Next, the redundancy existing in the descriptors data matrix was reduced by examining the descriptors' correlation with other descriptors and with the property of the molecules. Collinear descriptors (i.e. R > 0.9) were detected and the one presenting the highest correlation with the property was retained. The other collinear descriptors were removed from the data matrix.

2.3. Variable selection using replacement method (RM)

Theoretical researchers have concentrated a rising concentration on finding the most effective tools for choosing the best molecular descriptors in QSPR studies. Hence, there are many methods for the selection of the best structural descriptors from a large collection of them. One of such approaches is the replacement method (RM). The replacement method is a fast convergent iterative algorithm that produces linear regression models with small *S* in a particularly little computer time [38–40]. The RM is provided to help identify the best combination of descriptors from a large pool of variables.

In this case, we select an optimum subset $\boldsymbol{d}_\text{m} = \{X_{\text{m}1}, X_{\text{m}2}, \ldots, X_{\text{m}d}\}$ of *d* descriptors from a large set $\boldsymbol{D} = \{X_1, X_2, \ldots, X_D\}$ of *D* ones ($d \ll D$) provided by some available commercial program, with a minimum standard deviation (*S*):

$$S = \frac{1}{(N-d-1)}\sum_{\text{i}=1}^{\text{n}} res_\text{i}^2 \tag{2}$$

where *N* is the number of molecules in the training set, and res_i is the residual for molecule i (difference between the experimental and predicted property). Notice that $S(\boldsymbol{d}_\text{n})$ is a distribution on a separate space of $D!\,/\,d!(D-d)!$ disordered points $\boldsymbol{d}_\text{n}$. The full search (FS) that includes calculating $S(\boldsymbol{d}_\text{n})$ on all those points always allows us to arrive at the global minimum of *S*. The search is computationally unaffordable, if *D* is sufficiently large. The RM involves the following three steps [41]:

(i) We select an original set of descriptors $\boldsymbol{d}_k$ at random, replace one of the descriptors, say X_{ki}, with all the remaining $D - d$ descriptors, one by one, and retain the set with the smallest value of *S*. This is one step of the procedure.

(ii) In the consequent set, we select the descriptor with the greatest standard deviation in its coefficient and then substitute all the remaining $D - d$ descriptors in one-by-one method for it. We repeat this process until the set remains unchanged. In each cycle, we do not regulate the descriptor optimized in the previous cycle.
Therefore, we achieve the candidate $\boldsymbol{d}_\text{m}^{(\text{i})}$ that is derived from the so-constructed path i. It is worth noting that if the replacement of the descriptor with the largest error by those in the pool does not reduce the value of *S*, then that particular descriptor is not replaced.

(iii) We perform the above process for all of the possible paths $\text{i} = 1, 2, \ldots, d$ and keep the point $\boldsymbol{d}_\text{m}$ with the lowest standard deviation: $\min S(\boldsymbol{d}_\text{m}^{(\text{i})})$.

2.4. Support vector machine (SVM)

Support vector machine (SVM) is a mathematical entity, an algorithm for maximizing a particular mathematical function with respect to a given collection of data. SVM is a new and very talented classification and regression method developed by Vapnik et al. [18]. A comprehensive explanation of the theory of SVM can be referred in several excellent books and tutorials [42,43]. SVMs are originally developed for classification problems; they can also be extended to solve nonlinear regression problems by the introduction of ε insensitive loss function.

The basic idea in support vector regression (SVR) is to plan the input data X into a higher dimensional feature space *F* through a nonlinear mapping ϕ, and then a linear regression problem is obtained and solved